COMPUTER SIMULATION OF ARTICULATORY ACTIVITY IN SPEECH PRODUCTION

P. Mermelstein
Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey

A computer facility for modeling sound production resulting from exciting a moving artIculatory system is described. The vocal tract is treated as an acoustic transmission line with variable cross-sectional area. The system allows on-line specification of the history of artIculatory states (vocal-tract shapes) and excitation parameters, and generates the resultant acoustic signal for immediate evaluation by the experimenter. Novel features include graphical as well as numerical control of articulatory configurations, time-motion display of articulatory data and resonant-frequency versus time displays for the resulting acoustic signal. The facility is used for the study of the dynamics of articulatory movement and for speech synthesis.

Introduction

In modeling the process of human speech production we recognize five distinct stages. First, the message is organized on the linguistic level and structured grammatically. Second, the message is expressed phonologically in terms of a sequence of discrete units, labelled phonemes. Stress and intonation marks can be considered special phonemes and assigned special codes. Third, the string of phonemes is converted to a set of continuous signals controlling the musculature of the various articulators. This results in a highly complex integrated movement sequence in which generally all the articulators, the lips, the tongue, the mandible, etc., participate. Finally, the physical interaction of the vibrating vocal cords and the moving articulatory structure produces a continuous acoustic signal perceived as speech.

We are particularly interested in modeling the fourth or the articulatory movement stage of the speech event. We consider the speech event as composed of a sequence of phonetic segments specified in terms of one or more vocal-tract shapes and the corresponding excitation parameters. The acoustic signal is generated by simulating the excitation mechanism as well as the articulatory trajectory. In contrast, most previous speech production systems consider generation of the speech signal from the acoustic point of view and are controlled by specifying the variations of vocal-tract resonant frequencies (formant frequencies) with time.[1,2] Modeling the process at the articulatory level can be expected to be simpler because the articulators respond to muscular forces with predictable changes in their position and rates of movement. On the other hand, simple rules for resonant frequency changes are not necessarily realizable by the anatomy of the vocal tract. In fact, simple articulatory changes can have complex acoustic effects. This report is primarily concerned with generating the acoustic signal corresponding to given changes in the shape of the vocal tract. Accomplishment of that task is a prerequisite to studying the movements actually executed by the articulatory system when producing a particular utterance.

Recent studies show that no phoneme, not even the vowels, exhibits a unique set of resonant frequencies in all its occurrences in connected speech.[3] Therefore, it is unreasonable to expect that a unique vocal-tract shape can be associated with all productions of that phoneme. Certain attributes or distinctive features of the excitation mechanism and of the articulatory state serve, however, to differentiate among the phonemes. Such features may Include presence or absence of labial closure, of periodic or turbulent excitation, coupling of the nasal passage to the main vocal tract, etc. While Incorporation of these features in the articulatory description is necessary to obtain an intelligible result, it is not sufficient. The human speech perception apparatus is easily confused by false cues that arise from unnatural articulations. Therefore, those parameters unspecified by the distinctive features of the phoneme must be specified with reference to the moving articulatory structure so that the result will correspond to the natural execution of the articulation. Contextual effects such as anticipation or inertia control these optional features, and one is interested in specifying in detail the nature of that control mechanism.

If the effects of context are expressed in rules for articulatory movement, only the essential features of the phonemes need be specified explicitly,

and the optional parameters will be determined automatically. The simulation system is intended to aid the postulation of such rules by serving as a convenient tool for the evaluation of their articulatory and acoustic effects. This report describes a facility for specifying a discrete sequence of excitation and articulatory states of the vocal tract, and the system simulates the articulatory movement between these states, thus producing a continuous acoustic signal. The segment corresponding to a particular phoneme is represented as a sequence of one or more such states, where the numerical description of the states may be functions of the adjacent phonemes. For example, sustained tongue tip closure in the absence of voicing is a feature of the consonant /t/, but the shape of the tongue hump for that production is controlled by the adjacent vowels. Aspiration, another context-dependent feature of the /t/, is achieved by optionally introducing states specifying the onset and termination of turbulent excitation. For the moment, all parameters of the excitation and vocal-tract shape must be specified explicitly, or they are assigned fixed default values.

The timing of articulatory movement can be controlled in open-loop or closed-loop modes. The closed-loop mode would require continuous position or velocity feedback; however, the rapidity of the articulatory movement does not allow for this. Hence we assume a ballistic-type movement based on the present and the target articulatory positions. The time interval for the movement is explicitly specified, and it is assumed that the forces generated by the musculature are adjusted accordingly. The time durations are in practice functions of the adjacent articulatory states and future models can be expected to compute them from the articulatory specifications. Where timing differences arise from causes other than the articulatory context, such as the duration difference in English vowels before voiced and unvoiced consonants, these cannot be handled at the articulatory level and would have to be introduced explicitly even where articulatory movement proceeds completely by rule.

The speech production apparatus is modeled as consisting of a periodic source of variable frequency and amplitude, the vocal cords, and/or a noise source, exciting a nonuniform acoustic transmission line continuously changing in shape. The shape, of course, controls the characteristic impedance at any point, and thereby the transfer function of the line. Articulatory states are represented in terms of sampled values of the vocal-tract cross-sectional area at points along the tract.

Henke has postulated a set of rules governing tongue, lip and mandible movements based on analyzing frames from high-speed X-ray motion pictures of the moving vocal tract. However, he did not evaluate the resulting speech signal aurally. Our data are derived from acoustic measurement of the vocal-tract cross-sectional area5 and admit the postulation of similar rulea This report is primarily concerned with the implementation of a facility for immediate aural evaluation of utterances generated with the aid of postulated rules. Comparison of the computed trajectories of the vocal-tract resonant frequencies as functions of time with those measured from natural articulatory events yields a quantitative difference measure that can be used to complement the subjective evaluation of the audio signal. In addition, the specified movements are displayed for evaluation as sagittal projections convenient for comparison with X-ray pictures.

Static electric analogs of the vocal tract6,7,8 have been used for an extensive period as tools in the study of how deformations in the vocal tract control the characteristics of the speech signal. Dynamic electric analogs9,10 have been used for the synthesis of a restricted number of syllables, but they suffered from an inability to control the transmission-line elements over wide ranges at high speeds. Kadokawa and Nakatall have simulated the dynamic behavior of the vocal tract by computing a time-dependent transfer function and thus determining the variations with time of the vocal-tract resonant frequencies. Coker[12] uses a time-varying parametric representation of the vocal tract to compute the changing resonant frequencies and thus control a formant synthesizer. The primary obstacle to the development of a general synthesis facility has been the unavailability of a suitably flexible control system that allows quick and convenient specification or change of the parameters controlling the tract configuration. The computer-simulated system not only eliminates limitations to rapid configuration changes over a wide range of cross-sectional area values, but also permits experiments designed to search for convenient control strategies.

The presence of the human listener in the sound generation - evaluation - modification - sound generation feedback loop is necessary because at present we are unable to assign effective objective criteria to the significant attributes of

the acoustic result. Naturalness is a subjective criterion, and therefore the presence of the listener is essentia]. The experimenter also acts as a visual observer, evaluating shape and resonant-frequency variations with time. We endeavor to facilitate the experimenter's task so that in the light of his evaluation he can change the data variables quickly and conveniently. This is accomplished by allowing the experimenter to control the simulation using measures familiar to him, for example, lip opening or excitation frequency. He may build up the data for an utterance item by item, evaluating the result each time. Alternately, he may modify data previously entered and hear the modified audio signal in less than a minute.

The Simulation Method

The simulation principles are based on those used by Kelly and Lochbaum[13] with considerable extensions and modifications. The vocal tract is represented 'as a non-uniform transmission line supporting longitudinal propagation only. Its characteristic impedance at all points is inversely proportional to the cross-sectional area in a plane perpendicular to the flow line. The tract is assumed hard-walled; the compliance of the soft tissues of the tract wall is ignored. The sampling frequency for the simulation, 24 kHz, is determined by the spacing of area samples, 1.5 cm, that is necessary to simulate the continuous tract up to a signal frequency of 6 kHz. Since the length of each section is considerably smaller than the wavelength even at this frequency, the effect of the distributed reflection of a section traversed by the wavefront in one sampling period can be represented by a lumped reflection coefficient centered on the section.

Figure 1 shows a block diagram of the vocal-tract model. Pressure samples propagate back and forth along the transmission line and suffer multiple internal reflections. Excitation may be periodic, simulating vocal-cord vibration, or turbulent corresponding to frication. For periodic excitation the samples of the excitation function are added to the samples of the reflected signal arriving at the glottis. Turbulent excitation may take place at the vocal cords or at any interior section. It is Implemented by means of a white-noise generator and associated source impedance placed in series with the line. For computational simplicity, the tract is assumed lossless within its interior. Simple digital filters act at the tract boundaries, the vocal cords and the lips, to approximate the frequency-dependent losses so that actually measured resonant-frequency bandwidths are properly matched. The glottal reflection and the labial radiation functions act essentially as high-pass filters; the labial reflection appears as a low-pass filter. The nasal passage is represented by an additional transmission line of fixed but nonuniform shape coupled to the main tract with the aid of a variable coupling parameter. It serves in the production of nasalized vowels and consonants. The system output is obtained by summing the outputs of digital filters approximating the free-space radiation characteristics at the lips and nostrils. Radiation through the lips is, of course, controlled by the variable lip area. If the opening is reduced, more of the energy arriving at the lips is reflected within the tract and less is radiated to the outside. Radiation through the tract wall is not explicitly included.

Articulatory movement is simulated by defining a sequence of articulatory states or targets and interpolating the position-dependent reflection coefficient values and oral-nasal coupling between them. Thus targets are not necessarily phonemic in nature but may be specified as frequently as desired in order to approximate natural articulatory movement as closely as possible. When the articulator positions are derived from a model of articulatory activity, the vocal-tract shapes computed therefrom are expected to act as the target states of the simulation program. Stress and intonation information are assumed embedded in the excitation frequency and amplitude parameters and will not be discussed here in detail.

Linear interpolation of the reflection-coefficient values implies that near sharp constrictions the area varies linearly with time, but where the area changes slightly or the reflections are small, the areas vary exponentially with time. The resonant-frequency curves resulting from this procedure resemble those observed in spectrograms more closely than the straight-line segments used in most formant synthesizers.

Implementation

The system is operational on a DDP-224 computer. All of the routines are written in FORTRAN, except for the most frequently executed loop of the wave propagation computations which was rewritten in assembly language. Among the noteworthy hardware features utilized is the use of input/output channels to feed the loudspeakers through the digital-to-analog converter from one memory module, while simultaneously reading the next record from tape or disk into a second module.

Cathode-ray tube display proceeds simultaneously with processor operations on an interrupt basis.

The execution time of the simulation is approximately 40 times real time. The sample propagation computations along the transmission line, although basically very simple, are the most time consuming. A special slave processor has been designed to carry out these operations under control of the general purpose computer. If implemented, it would cut the simulation time to four times real time. Immediate application to real-time speech synthesis is therefore limited and the system is considered primarily a research tool.

The excitation period serves as a time clock for the synthesis routine, and within each period parameters or parameter increments remain constant. Therefore an excitation period is specified even when the tract is excited by noise or not excited at all as in pauses within the utterance. This alllows a relatively rapid, free-running simulation for hundreds of sampling periods before a check is made for new parameter values.

Once the control data have been specified the simulation runs without interaction, and writes the speech output on disk or tape. Thus the time duration of speech material generated in any run is limited only by the control table storage limitations and the patience of the experimenter. Because the pressure distribution along the vocal tract at any time within the utterance cannot be simply recreated, a change in any one target parameter requires the regeneration of the complete utterance.

Control and Use of the Simulation Facility

The control parameters for a particular synthesis task are specified in terms of two parameter lists. The first is a list of articulatory states defined in terms of sets of 13 area values. The second list is a time sequence of targets, nasal coupling values, target Interval durations, excitation frequency values, excitation amplitude values, whether periodic or random excitation and if the latter, the place of excitation. Targets are specified by using character-string labels as pointers to entries in the table of articulatory states. Excitation parameters are conveniently separated from specification of the articulatory configurations and each is easily varied independently of the other. The same articulatory configuration may be repeatedly called by its label in the second list without having to specify it again in numerical terms. The program supplies default values for most parameters, generally unchanged from the previous specification so that only changing parameters need be explicitly specified. Targets may be appended, inserted or deleted from the list conveniently and quickly. A typical sentence, the example discussed later, which consists of 19 phonemes required a sequence of 27 articulatory targets for acceptable synthesis.

The basic mode of data entry is through the on-line typewriter. Once a parameter list satisfactory to the experimenter has been entered, the program generates the output signal and records it digitally on tape or disk. Because the simulation runs slower than real time, the output signal is converted to analog form and played back only after its generation has been completed.

The above method of operation, though perfectly feasible, is not terribly convenient from the point of view of an experimenter not used to numerical specification of articulatory data. To visualize the articulatory configurations and movements between them, a time-motion display of the articulatory trajectory is projected at a rate proportional to the real-time movement rates specified. Any particular configuration may be selected for static display, graphically modified with the aid of a visible pointer, and returned to its slot in the target history of the utterance.

Another display that has found wide application to evaluate the course of articulatory events is *one* where the sequence of vocal-tract resonant frequencies are displayed on a time base. To ease the computational load, the resonances are computed from the stored shapes!3 rather than the generated signal. This display resembles a conventional spectrogram in appearance. Based on his wide experience with spectrograms, an experienced speech scientist can rapidly pin-point situations where the acoustic effects of the specified articulatory trajectories deviate appreciably from normal patterns. The visual Information he obtains thus complements his aural evaluation of the generated utterance.

Let us now consider in detail simulation of the production of the sentence, "This is a story about a man." This sentence was generated through trial and error procedures by modifying shapes derived from X-ray pictures and referring to the natural utterance for timing values. It demonstrates the quality of output available through careful work with the system. No explicit articulatory rules were used in this example. Instead the

experimenter's accumulated knowledge of articulatory dynamics and the acoustics of speech production were utilized. The automation of this process so that one will not have to resort to trial and error for every different sentence generated is the next task at hand.

In Figure 2 the spectrograms of the synthetic and natural utterances are compared. We transcribe that sentence in terms of phonemic symbols ais lz a 'storl ''baUt » 'main/. The apostrophes indicate stress marks which are considered in specifying the excitation frequency and target interval duration data. The excitation frequency function for the sentence is obtained by modifying the function determined from the intonation pattern. The excitation frequency is increased prior to and decreased relatively more rapidly after stress marks. For the voiced fricatives /$/ (th) and /z/ we supply, in addition to periodic excitation, turbulent excitation at points 3 cm and 1.5 cm respectively behind the front end of the tract. In each case the vocal-tract shape is relatively constricted at that point. The unvoiced fricative /s/ is excited by noise only at the same point and using the same vocal-tract shape as /z/. The vowels /*%, /«/, /o/ and /at/ are periodically excited and have their articulatory configurations shaped by modifying the configuration appropriate for the isolated vowel in accordance with the articulatory influences of the neighboring consonants. Tn turn, the vowels are used to control the variable articulatory characteristics of the consonants. For example, the articulation of the initial /s/ is the same as that of the /%/ except where the tongue tip is raised to produce the constriction. The glide /r/ is vowel-like in excitation and retroflexion of the tongue is not applied. The diphthong /a©/ is generated by linear reflection-coefficient interpolation between bounding target values that are somewhat modified versions of the isolated independent vowels /a/ and /u/. The unvoiced stop /t/ has a transition part to an articulation exhibiting almost complete constriction at a point 3 cm from the front. This constricted articulatory state is sustained for a 21 msec interval before release. On release a short burst of noise is introduced to simulate affrication or turbulence; thereafter the amplitude of the periodic excitation is linearly increased from a zero value. The voiced stop /b/ has the articulation changing from that of the previous vowel /*/ to that of the initial part of the following diphthong /aur/ while the lips are kept almost completely constricted. A very small but finite lip opening is helpful in simulating the acoustic effects of energy radiated through the walls of the mouth. The nasal consonants /m/ and /n/ are produced with the oral tract completely constricted at the lips and 3 cm behind them, respectively, but with a high value of nasal coupling. The coupling is reduced but not eliminated for the intervening ea/ vowel which is thus made nasa-1ized.

One interesting result observed through trial and error experiments is that naturalness is enhanced by avoiding sustained articulations - those where two adjacement articulatory states are specified as identical. Except for the vowels in the stressed syllables and the fricatives, the articulations are continuously changing. This further invalidates the idea that speech can be produced by concatenating different stationary signals.

Conclusions

The simulation program produces intelligible speech if care if taken to modify articulatory shapes, time durations, and excitat ion frequency values to fit the contextual environment on a level at least as large as the sentence. Data appropriate for the production of an isolated word but not for the sentence in which it is embedded may make not only that word, but the complete sentence unintelligible. The quality of the material generated so far can be used as a goal for the quality to be achieved by implementing rules governing the movements of the individual articulators.

The program uses an excitation function that is stored and independent of the vocal-tract shape. The articulation, however, is known to affect both the wave-shape and the excitation frequency.15 The fixed wave-shape independent of frequency is but a rough approximation to the real situation. Detailed simulation of vocal-cord movement would, however, increase the execution time of the program by at least a factor of two. Because these factors are not central to the aspects of speech generation under consideration at the moment, they have not been included in the present program.

The vocal tract is assumed lossless in its interior, therefore, the resulting resonance bandwidths are fixed functions of frequency and independent of the articulation. The losses in general do depend on the surface area of the tract wall and are thus affected by the shape of the tract in the plane perpendicular to the flow directions. The contribution of this factor to lack of naturalness has not been precisely evaluated.

Articulations of /l/ and /r/, in particular, often exhibit sidebranches formed by the tongue and the hard palate that cannot be modeled precisely when one treats the vocal tract as one continuous tube. Our approximations are therefore expected to lead to differences in the acoustic result and must be recognized as limitations of the model.

A program that requires explicit specification of all articulatory and excitation parameters suited to the particular context cannot be said to exhibit extensive artificial intelligence. Instead, the simulation program must be viewed as a framework within which statements regarding articulatory movements can be expressed and evaluated. The framework reported is based on the consideration that utterances are organized on the articulatory level, and that organization can be studied most effectively at the same level.

Our results up to this time concern organization at the lowest level. Here articulatory parameters are treated as independent, and interactions between them, except as reflected on the acoustics level, are ignored. For example, the nasal coupling parameter is used to control nasalization of vowels adjacent to a nasal consonant at the same time as the tongue shape is changing as required for the production of those vowels. Thus the adjacent vowels are perceived as nasalized. The production of the same effect on the acoustic level is much more complex. This example substantiates the basic motivation for speech generation through articulatory control, that this direction of attack can unravel the complex organization of the individual phonemes into the integrated utterance.

References

1. Holmes, J. N., Mattingly, I. G. and Shearme, J. N., "Speech Synthesis by Rule." Language and Speech, *7,* 127 (1964).

2. Rabiner, L. R., "Speech Synthesis by Rule: An Acoustic Domain Approach, Bell System Tech. Journal 47, 17 (1968).

3. Lindblom, B., "Spectrographs Study of Vowel Reduction," J. Acoust. Soc. Am., 35, 1773,(1963).

4. Henke, W., "Dynamic Articulatory Model for Speech Production Using Computer Simulation," Ph.D. Thesis, Mass. Inst, of Tech., Cambridge, Mass., 1966.

5. Schroeder, M. R., "Determination of the Geometry of the Human Vocal Tracts by Acoustic Measurements," J. Acoust. Soc. Am., 41, 1002 (1967).

6. Dunn, H. K-, "The Calculation of Vowel Resonances and an Electrical Vocal Tract," J. Acoust. Soc. Am., 22, 740 (1950).

7. Stevens. K. N., Kasowski, S. arid Pant, C. G. M., "An Electric Analog of the Vocal Tract," J. Acoust. Soc. Am., 25, 734 (1953).

3. Fant, C. G. M., Speech Communication Research, Ing. Veitenshaps Akad., (Stockholm), 24, 331 (1953)

9. Rosen, G., "Dynamic Analog Speech Synthesizer," J. Acoust. Soc. Am., 30, 20 (1953).

10. Hecker, M. H. L., "Studies of Nasal Consonants With an Articulatory Speech Synthesizer," J. Acoust. Soc. Am., 34, 179 (1962).

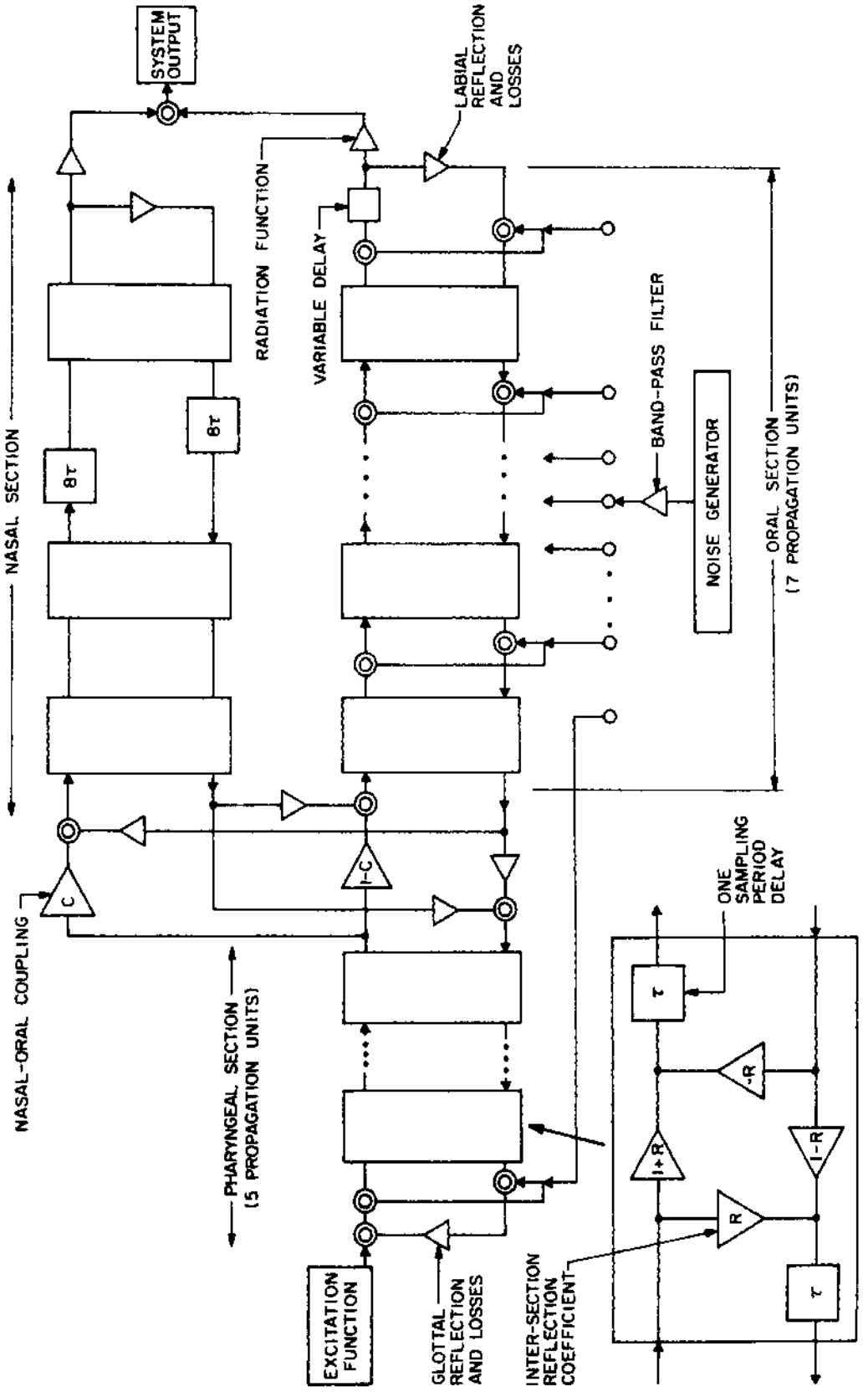
11. Kadokawa, Y. and Nakata, K., "Analysis of Speech by Vocal-Tract Configuration," J. Radio Res. Laboratories, Japan, 11, 99 (1964).

12. Coker, C. H., "Speech Synthesis with a Parametric Articulatory Model," paper A-4, Speech Symposium, Kyoto, Japan, 1968.

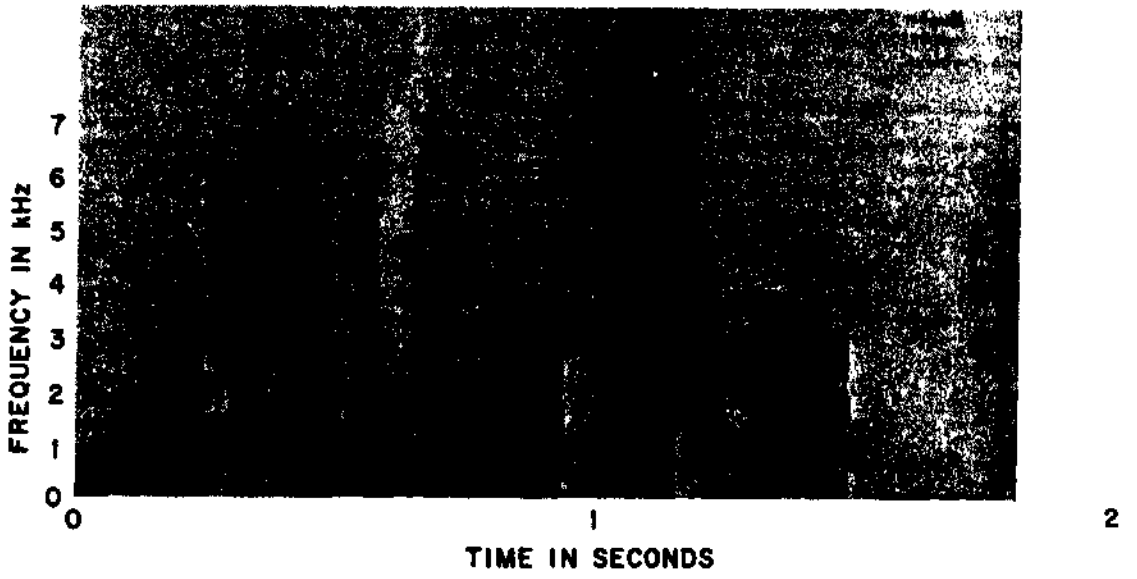
13. Kelly, J. L., Jr. and Lochbaum, C. C, "Speech Synthesis," Proceedings of the Fourth International Congress on Acoustics, Copenhagen, 1962.

14. Mermelstein, P., "Determination of the Vocal-Tract Shape from Measured Formant Frequencies," J. Acoust. Soc. Am., 41, 1283 (1967).

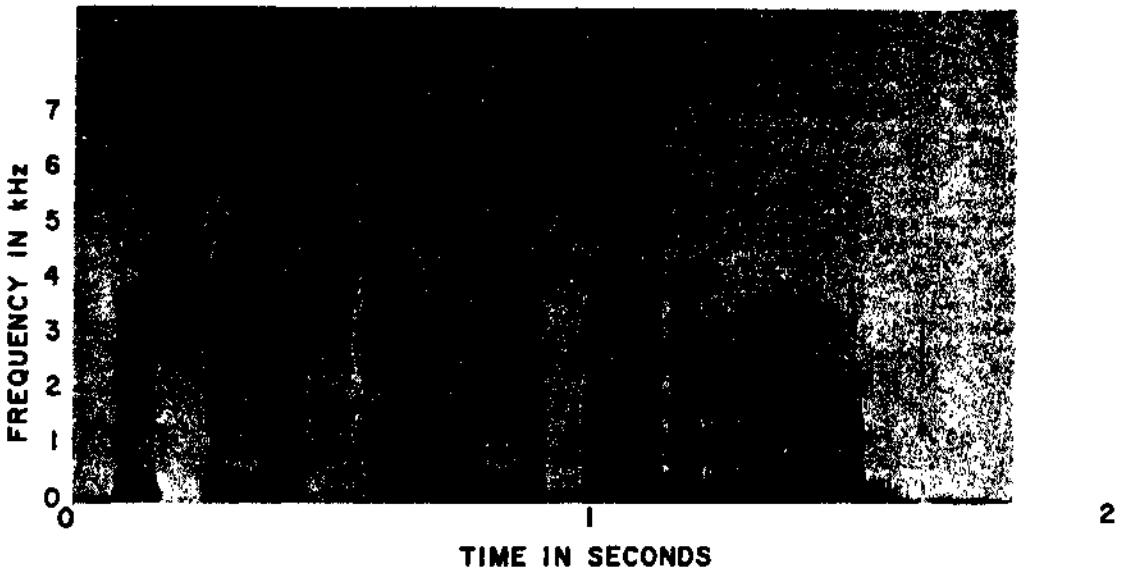
15. Flanagan, J. L. and Landgraf, L. L., "Self-Oscillating Source for Vocal-Tract Synthesizers," Trans. IEEE AU-16, 57-64 (1968).



1. Block diagram of the simulator system.



MACHINE GENERATED SENTENCE



NATURAL SENTENCE

2. Spectrograms of machine generated and natural versions of "This is a story about a man."